International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2017 IJSRCSEIT | Volume 2 | Issue 3 | ISSN : 2456-3307

An Enhanced Technique for Identifying Cancer Biomarkers from Microarray Data Using Hybrid Feature Selection Technique

Dr. K. Kalaivani[1], S. Senthil Kumar [2]
[1]Associate Professor, Department of Computer Applications (PG),
[2]Assistant Professor, Department of Commerce with Computer Applications,
[2]Dr.SNS Rajalakshmi College of Arts and Science (Autonomous), Coimbatore, Tamil Nadu, India

ABSTRACT

Cancer is one of the fearful diseases found in majority of the living organism, and is one of the demanding focuses for scientists from 20th century. Cancer research is one of the major research areas in the medical field. There were bunch of proposals from a variety of establishers and detailed picture examination was still under processing. Fundamentally Cancer is described as an abnormal, uncontrolled growth that may demolish and invade neighbouring healthy body tissues or elsewhere in the body. Living organisms like animals and plants consist of cells. The simplest organisms contain only a single cell. The human body consists of billions of cells; majority of the cells include a restricted life-span and require being replaced cyclic manner. Every cell is competent of duplicating themselves. Millions of cell divisions and replications happen daily in the body and it is shocking that the process happens so entirely and most of the time every cell division needs replication of the 40 volumes of genetic coding. On rare situation there is some fault in a division and a rogue, potentially malignant cell arises. The immune system appears to distinguish such occurrences and is normally proficient of removing the abnormal cells before they have an opportunity to proliferate. On the odd occasion, there is a collapse of the mechanism and a potentially malignant cell survives, replicates and cancer is the consequence.

Keywords: Cancer Biomarkers, Feature Selection Technique

I. INTRODUCTION

With more than 1,300 persons succumbing to cancer every day, it has become one of the major causes of death occurring in the country due to communicable and life-style ailments, followed by tuberculosis.

As per data of the National Cancer Registry Programme of the Indian Council of Medical Research (ICMR), the estimated mortality rate due to cancer saw an increase of six percent approximately between 2012 and 2014.

"There has been close to 5 lakh deaths due to cancer in the country in 2014," said a senior Health Ministry official.

Total of 4,91,598 people died in 2014 out of 28,20,179 cases, while in 2013 it was 4,78,180 deaths out of 29,34,314 cases reported and in 2012, around 4,65,169 people lost their lives due to the disease when the number of cases stood at 30,16,628.[23]

The computer era poses a challenge for many healthcare professionals and Information technology (IT) has broad impact to medicine. It combines statistics, visualization, machine learning and other intelligent techniques in order to analyze large medical databases. The idea of this research is to deliver an interdisciplinary course focusing on soft computing methods in medicine. IT provides key methods of thinking and problem solving which fundamentally affect all of healthcare. However, medical informatics is a broad interdisciplinary field, and there is a need to promote the kind of multidisciplinary thinking required in this field.

II. CANCER BIOMARKERS

In the recent years, knowledge about cancer biomarkers has increased tremendously providing great opportunities for improving the management of cancer patients by enhancing the efficiency of detection and efficacy of treatment. Recent technological advancement has enabled the examination of many potential biomarkers and renewed interest in developing new biomarkers. Biomarkers of cancer includes a broad range of biochemical entities, such as nucleic acids, proteins, sugars, lipids, and small metabolites, cytogenetic and cytokinetic parameters as well as whole tumour cells found in the body fluid.

Biomarkers are therefore invaluable tools for cancer detection, diagnosis, patient prognosis and treatment selection. These can also be used to localize the tumour and determine its stage, subtype, and response to therapy. Biomarkers are subject to dynamic modulation, and are expected to enhance our understanding of drug metabolism, drug action, efficacy, and safety. Advanced clinical practice in certain malignancy have effectively used tumour and immune cells where it served as a good biomarker of prognosis, while its utility in other cancers are under evaluation at the present time.

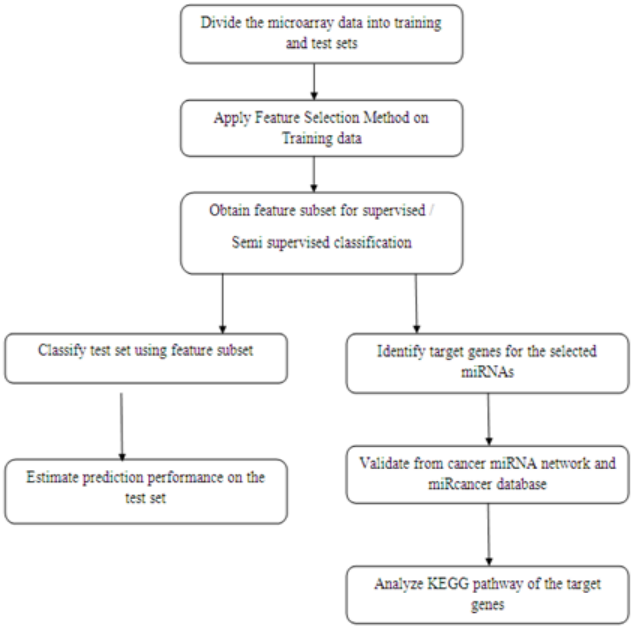


Figure 1. Procedure for Cancer Biomarkers Identification

III. MICROARRAY TECHNOLOGY

The recent advent of microarray technology has allowed the simultaneous monitoring of thousands of genes, which motivated the development in cancer classification using gene expression data (Slonim et al.,). Though still in its early stages of development, results obtained so far seemed promising.

Microarray technology is a developing technology used to study the expression of many genes at once in single experiment an a small chip. It is also called as Gene chip or bio chip which is used to analyze gene expression.

Microarray technology has the potential to address many interesting questions in genetics by revealing patterns of expression for genes and classifying samples (such as tumour samples) based on such patterns. However, basic questions about microarray data persist without satisfactory answers. The simplest microarray experiment studies the variation in gene expression across the categories of a single factor, such as tissue types, strains of mice, or drug treatments. The purpose of such an experiment is to identify differences in gene expression among the varieties.

The abundance of data resulting from a single microarray assay has sometimes fostered a distorted view that microarray data can be collected in a relatively unplanned manner. The naive expectation is that the sheer volume of data generated will suffice for algorithmically determining important and unanticipated patterns in the data. This is not a good plan for using microarray technology; microarray studies require careful planning and development of analysis strategies. There is a tension in biology research between ‘‘hypothesis driven’’ research and ‘‘descriptive’’ research. Descriptive research is often suspect because it is not designed to answer specific questions and therefore the usefulness of the data collected may be questioned. Microarrays have been seen as a tool for descriptive research because they provide a survey of gene expression rather than a focus on mechanistic aspects of the workings of a small number of genes.

Microarrays generally are notbest suited for testing gene-specific mechanistic hypotheses because other more sensitive assays are available for measuring expression of a specific gene. Nevertheless, most effective microarray-based studies have a clear objective and answer well-defined questions, although generally not gene-specific mechanistic questions. Clear identification of the objective of a microarray

study is important for designing the study and constructing an appropriate analysis strategy.

Another type of microarray study involves the identification of novel subtypes of specimens within a population. This objective is based on the idea that important biological differences among specimens that are clinically and morphologically similar may be discernible at the molecular level.

Machine learning is a bough of Artificial Intelligence (AI) that uses a variety of statistical, probabilistic and optimization systems that permits computers to "learn" from past examples and to detect hard-to-discern patterns from large, noisy or complex data sets. Therefore, machine learning is frequently used in cancer diagnosis and detection.

IV. TYPES OF MICROARRAY

Microarrays can be broadly classified according to at least three criteria:

- Length of the probes
- Manufacturing method
- Number of samples that can be simultaneously profile on one array**.**

Many microarray studies in cancer have the objective of developing a taxonomy of cancers that originate in a given organ site in order to identify subclasses of tumors that are biologically homogeneous and whose expression profiles either reflects different cells of origin or other differences in disease pathogenesis. These studies may uncover biological features of the disease that pave the way for development of improved treatments by identification of molecular targets for therapy.

V. APPLICATIONS OF MICROARRAY

Microarray can be a boon to researchers as it provides a platform for simultaneous testing of a large set of genetic samples. An important application of microarray gene expression data in functional genomics is to classify samples according to their gene expression profiles.

- Gene discovery
- Gene selection
- Disease diagnosis
- Drug discovery

Microarray data pre-processing is a very important stage for classification. One crucial step of Microarray data pre-processing is gene selection. A good gene selection method cannot only increase the accuracy of classification by eliminating the irrelevant genes from Microarray data, but also speed up the classification process by reducing the Microarray data size.

The *existing approach* – the combination Inductive SVM and Transductive Support Vector Machine (TSVM) suffers from a low signal-to-noise ratio, which causes instability in gene signatures. Hence, to improve prediction accuracy, efficient dimensionality reduction techniques need to be explored. Furthermore, inadequate observations of gene / miRNA data result in poor performance of the traditional supervised methods.

VI. OBJECTIVES OF THE RESEARCH

- Different combination of two or more feature selection (Hybrid Feature Selection Technique) methods needs to be investigated to obtain more biologically relevant genetic signatures.
- To find cancer related genes and hence to develop better diagnostic methods.
- To extract top ranked genes using Fuzzy preference based Rough Set.
- To remove / extract irrelevant or redundant genes,this is helpful for biologists to find cancer related genes.

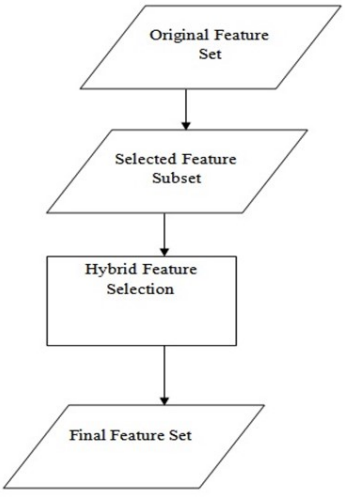


Figure 1. Hybrid Feature Selection Procedure

VII. PROPOSED METHODOLOGY:

The objective of Hybrid Feature Selection `is to extract a subset of relevant features which is useful for model generation. Many mining algorithms do not perform well with large number of features. Hence, to improve prediction accuracy, efficient dimensionality reduction techniques need to be explored.

FUZZY SET:

A Fuzzy Set consists of linguistic variables where values are words and not numerical. Linguistic variable have fuzzy margins and can overlap each other. The transition from one value to another is gradual and each value is given a membership function which represents the degree to which it belongs to that value.

A Fuzzy Set is any set that allows its members to have different grades of membership in the interval [0-1]. In Fuzzy applications, the non-numeric linguistic variables are often used to facilitate the expression of rules and facts. However, Fuzzy was used in more useful applications. Fuzzy Logic has proved to be particularly useful in Expert Systems and other Artificial Intelligence Applications.

FEATURE SELECTION TECHNIQUE:

Feature Selection (FS) aims to decrease the dimensionality of large scale data sets without losing useful information. Feature selection offers a number of advantages, including more powerful classification models by eliminating irrelevant or noisy features, more compact and faster models by constructing them using only a small subset of the original set of features, and the ability to focus on a subset of relevant features, which can be used for the discovery of new knowledge. This work focuses on the use of feature selection techniques for biomarker discovery from microarray data. It is a common practice for a domain expert to start validating the biomarkers selected by the feature selection algorithm.

The same feature selection technique may produce drastically different results depending on the chosen setting of the parameters of the method. To make matters even worse, many of the current datasets are described by a number of features that generally exceed the number of available training samples.
Selection of relevant genes for a given pathology on different sub-samplings of the patients should produce nearly the same results since the biological process generating the data is assumed to be largely common for all patients, at least without confounding factors.

DIFFERENT FEATURE SELECTION TECHNIQUES:

- Kernelized Fuzzy Rough Set for Feature Selection.
- Feature selection using Fuzzy preference based Rough Set .
- Consistency Based Feature Selection.
- Signal -to-Noise Ratio for Feature Selection

ADVANTAGES OF FEATURE SELECTION:

- First, dimension reduction is employed to reduce the computational cost.
- Second, reduction of noises is performed to improve classification accuracy.
- Finally, extraction of more interpretable features or characteristics that can be helpful to identify and monitor the target diseases.

VIII. PUBLICLY AVAILABLE MICROARRAY DATA SETS

Since classification is a typical and fundamental issue in diagnostic and prognostic prediction of cancer, different combinations of methods are studied using the four datasets.

- ***Small Round Blood Cell Tumors (SRBCT):***The Small round blood cell tumors are four different childhood tumors named so because of their similar appearance on routine histology. The number of samples is 83 and total number of genes is 2308. They include Ewings sarcoma (EWS) (29 samples), neuroblastoma (NB) (18 samples), Burkitt's lymphoma (BL) (11 samples) and rhabdomyosarcoma (RMS) (25 samples).
- ***Diffuse Large B-Cell Lymphomas (DLBCL):***Diffuse large B-cell lymphomas and follicular lymphomas are two B-cell lineage malignancies that have very different clinical presentations, natural histories and response to therapy. The dataset contains 77 samples and 7070

genes. The subtypes are diffuse large B-cell lymphomas (DLBCL) (58 samples) and follicular lymphoma (FL) (19 samples).

- ***Leukemia:***Leukemia is an affymetrix high-density oligonucleotide array that contains 5147 genes and 72 samples from two classes of leukemia: 47 acute lymphoblastic leukemia (ALL) and 25 acute myeloid leukemia (AML).
- ***MicroRNA Dataset:*** We have downloaded a publicly available miRNA expression dataset from the website: http://www.broad.mit.edu/cancer/pub/miGCMThe dataset contains 217 mammalian miRNAs from different cancer types. From this, we have selected six datasets consisting of the samples from colon, kidney, prostate, uterus, lung and breast. Each dataset is presented by all the 217 miRNAs.

Each sample vector of the datasets is normalized to have mean 0 and variance 1. The resulting single dataset contains two classes of samples, one representing all the normal samples with 32 examples and another representing tumor samples having 43 examples. The dataset is first randomized and then partitioned into training (38 samples) and test set (37 unlabeled samples). While dividing into training and test sets, it is ensured that both training and test sets contain at least one sample from normal and malignant samples of each of the tissue types. Feature selection algorithms are applied on the training set to extract informative miRNAs.

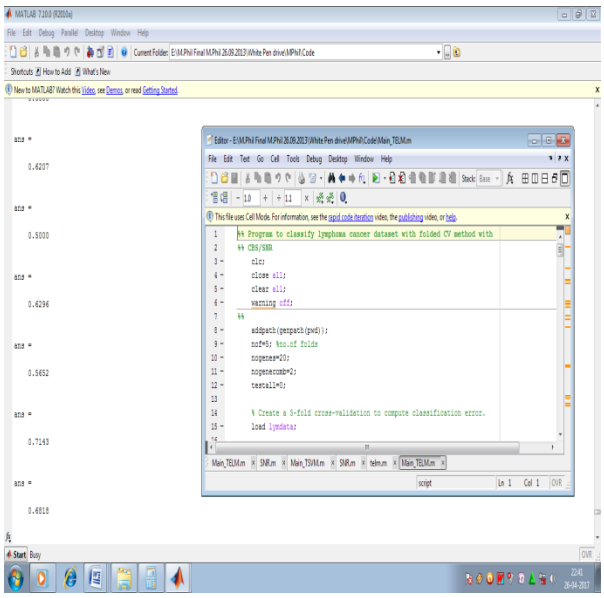
TABLE 1. The number of normal and tumor samples present in each tissue type.

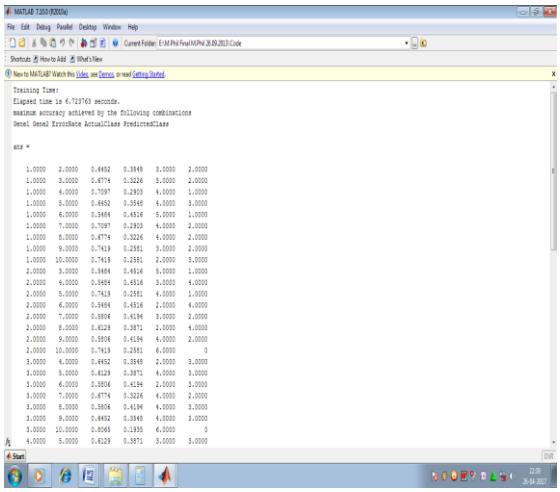
Tissue	Normal Samples	Tumor Samples	Total
Colon	5	8	13
Kidney	3	4	7
Prostate	8	4	12
Uterus	9	8	17
Lung	4	5	9
Breast	3	2	5
Total	32	31	63

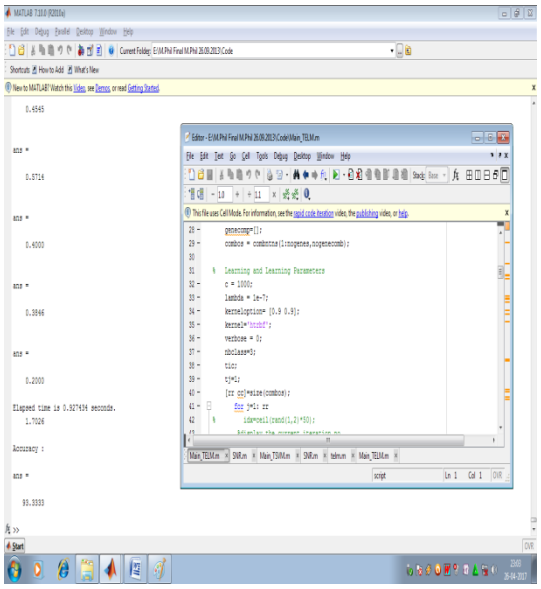
TABLE 2. Over all accuracies and Standard deviation using T-Statistic

Data set	Test Set	Training Set	KFRS + TSVM + KNN	KFRS + ISVM + KNN	FPRS + TSVM + KNN	FPRS + ISVM + NB	CBFS + TSVM + NB	CBFS + ISVM + NB
SRBCT	38	7	96.45	93.65	95.46	89.21	91.12	83.56
		12	98.49	93.28	97.32	91.75	92.05	86.29
		17	98.06	93.33	97.38	93.65	96.21	91.21
DLBCL	34	7	97.45	92.55	96.27	87.24	90.52	86.28
		12	97.98	92.79	98.34	89.52	91.87	87.25
		17	98.90	94.81	94.71	91.25	93.45	88.25
Leukemia	32	7	99.99	95.69	95.52	93.26	90.12	83.12
		12	96.07	96.03	96.34	94.12	93.54	87.25
		17	98.32	97.04	97.21	80.41	92.62	77.56
miRNA	31	7	97.28	94.06	86.58	89.23	82.65	78.21
		12	97.62	95.11	94.07	87.29	82.42	81.25
		17	98.08	97.05	95.11	89.25	85.71	84.59

IX. RESULTS AND DISCUSSION







KERNELIZED FUZZY ROUGH SET FOR FEATURE SELECTION:

High level of similarity between kernel methods and rough sets can be obtained using kernel matrix as a relation [13]. Kernel matrices could serve as fuzzy relation matrices in fuzzy rough sets. Taking this into account, a bridge between rough sets and kernel methods with the relational matrices was formed [13]. Kernel functions are used to derive fuzzy relations for rough sets based data analysis. In this study, Gaussian kernel approximation has been used to construct a fuzzy rough set model, where sample spaces are granulated into fuzzy information granules in terms of fuzzy T -equivalence relations computed with Gaussian kernel.

X. CONSISTENCY BASED FEATURE SELECTION

Dash and Liu introduced consistency function that attempts to maximize the class separability without deteriorating the distinguishing power of the original features. Consistency measure is computed using the properties of rough sets. Rough sets provide an effective tool which deals with the inconsistency and incomplete information. This measure attempts to find a minimum number of features that separate classes as consistently as the full set of features can. In classification, it is used to select a subset of original features which is relevant for increasing accuracy and performance, while reducing cost in data acquisition. When a classification problem is defined by features, the number of features can be very large, many of which are likely to be redundant. Therefore, a feature selection criterion is defined to select relevant features. Class separability constraint is usually employed as one of the basic selection criteria. Consistency measure can be used as a selection criterion that heavily depends on class information and aims to keep the discriminatory power of the actual features. This measure is defined by inconsistency rate and its method of computation can be found in [22] .

XI. CONCLUSION

Gene combinations classify different cancer types, lead to a better understanding of genetic signatures and cancer biomarkers to improve accuracy. Integration of other sources of information could be important to enhance clinical / translational research. For example, model development where both clinical variables and gene / miRNA expression can be combined to improve prediction power Hybrid Feature Selection Technique obtain more biologically relevant genetic signatures.

XII. REFERENCES

[1]. E. Berezikov, E. Cuppen, and R. H. A. Plasterk, “Approaches to microRNA discovery," Nature Genet., vol.38, pp. S2_S7, May 2006.

[2]. S. Bandyopadhyay, R. Mitra, U. Maulik, and M. Q. Zhang, ``Development of the human cancer microRNA network," BMC Silence, vol. 1, no. 1, p. 6,2010.

[3]. S. Bandyopadhyay, A. Mukhopadhyay, and U. Maulik, ``An improved algorithm for clustering gene expression data," Bioinformatics, vol. 23, no. 21, pp. 2859_2865, 2007.

[4]. U. Maulik and A. Mukhopadhyay, ``Simulated annealing based automatic fuzzy clustering combined with ANN classi_cation for analyzing microarray data," Comput. Oper. Res., vol. 37, no. 8, pp. 1369_1380, Aug. 2010.

[5]. A. Mukhopadhyay and U. Maulik, ``Towards improving fuzzy clustering using support vector machine: Application to gene expression data," Pattern Recognit., vol. 42, no. 11, pp. 2744_2763, Nov. 2009.

[6]. U. Maulik, ``Analysis of gene microarray data in a soft computing framework," Appl. Soft Comput., vol. 11, no. 6, pp. 4152_4160, Sep. 2011.

[7]. U. Maulik, S. Bandyopadhyay, and A. Mukhopadhyay, Multiobjective Genetic Algorithms for Clustering: Applications in Data Mining and Bioinformatics. New York, NY, USA: Springer-Verlag, 2011.
[8]. S. Bandyopadhyay, U. Maulik, and J. T.Wang, Analysis of Biological Data: A Soft Computing Approach. Singapore: World Scienti_c, 2007.
[9]. L.-K. Luo, D.-F. Huang, L.-J. Ye, Q.-F. Zhou, G.-F. Shao, and H. Peng, ``Improving the computational efficiency of recursive cluster elimination for gene selection," IEEE Trans. Comput. Biol. Bioinformat., vol. 8, no. 1,pp. 122_129, Jan./Feb. 2011.
[10]. S. Ramaswamy et al., ``Multiclass cancer diagnosis using tumor gene expression signatures," Proc. Nat. Acad. Sci., vol. 98, no. 26, pp. 15149_15154, 2001.
[11]. U. Maulik and D. Chakraborty, ``Fuzzy preference based feature selection and semisupervised SVM for cancer classi_cation," IEEE Trans.Nanobiosci., vol. 13, no. 2, pp. 152_160, Jun. 2014.
[12]. D. C. Koestler et al., ``Semi-supervised recursively partitioned mixture models for identifying cancer subtypes," Bioinformatics, vol. 26, no. 20, pp. 2578_2585, 2010.
[13]. Q. Hu, D. Yu, W. Pedrycz, and D. Chen, ``Kernelized fuzzy rough sets and their applications," IEEE Trans. Knowl. Data Eng., vol. 23, no. 11,pp. 1649_1667, Nov. 2011.
[14]. Q. Hu, L. Zhang, D. Chen, W. Pedrycz, and D. Yu. Gaussian Kernel Based Fuzzy Rough Sets: Model, Uncertainty Measures and Applications.Online]. Available: http://www4.comp.polyu.edu.hk/
[15]. Q. Hu, D. Yu, and M. Guo, ``Fuzzy preference based rough sets," Inf. Sci., vol. 180, no. 10, pp. 2003_2022, 2010.
[16]. M. Dash and H. Liu, ``Consistency-based search in feature selection," Artificial Intelligence., vol. 151, nos. 1_2, pp. 155_176, Dec. 2003.
[17]. T. R. Golub et al., ``Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring," Science, vol. 286, no. 5439, pp. 531_537, 1999.
[18]. H. Peng, F. Long, and C. Ding,``Feature selection based on mutual information criteria of max dependency,max-relevance, and min-redundancy," IEEE Trans. Pattern Anal. Mach. Intell., vol. 27, no. 8, pp. 1226_1238, Aug. 2005.
[19]. Online]. Available: http://www.biolab.si/supp/bi-cancer/projections/
[20]. J. Lu et al., "MicroRNA expression profiles classify human cancers", Nature, vol. 435, no. 7043, pp. 834_838, Jun. 2005.
[21]. Debasis Chakraborty and Ujjwal Maulik, (Senior Member, IEEE), Identifying Cancer Biomarkers From Microarray Data Using Feature Selection and Semi-supervised Learning ,IEEE Journal of Translational Engineering in Health and Medicine, vol.2, 2014, pp.1-11, Dec 2014
[22]. U. Maulik, A. Mukhopadhyay, and D. Chakraborty, ``Gene-expression based cancer subtypes prediction through feature selection and transductive SVM," IEEE Trans. Biomed. Eng., vol. 60, no. 4, pp. 1111_1117, Apr. 2013.
[23]. M. Dash and H. Liu, ``Consistency-based search in feature selection," Artif.Intell., vol. 151, nos. 1_2, pp. 155_176, Dec. 2003.
[24]. http://articles.economictimes.indiatimes.com